Джастин Хатчинс

ЯЗЫК ОБМАНА

КАК ИИ НОВОГО ПОКОЛЕНИЯ СТАНОВИТСЯ ОРУЖИЕМ

НА ГРАНИ ДОВЕРИЯ: ЧЕЛОВЕК И ЕГО УМНЫЕ МАШИНЫ

Чем дальше, тем больше мы доверяем умным помощникам — от поисковых и рекомендательных алгоритмов до устройств, установленных в наших домах, и инструментов для решения рабочих задач. Мы отмахиваемся от паникеров, опасающихся восстания машин, и увлеченно общаемся с ботами обо всем на свете — от бытовых вопросов до рассуждений о смысле жизни… Но всегда ли мы понимаем, что растущее доверие к машинам без должного контроля действительно может обернуться — и уже оборачивается — против нас?

Эксперт в вопросах кибербезопасности и социальной психологии Джастин Хатчинс предлагает читателю не просто обзор современных технологий, а комплексный взгляд на логику развития ИИ и структурированное предупреждение о рисках, которые нужно учитывать уже сегодня.

Эта книга — обязательное чтение для всех, кто интересуется ИИ, кибербезопасностью, философией сознания и будущим цифрового общества. Хатчинс объясняет, как работает социальная инженерия в исполнении ИИ, какие уязвимости существуют в архитектуре моделей-трансформеров, почему тест Тьюринга нерелевантен для современных умных помощников и что люди могут и должны сделать в сфере контроля над ИИ, пока не стало слишком поздно.

Читайте это саммари, чтобы отличать реальность от фантастики даже там, где они практически слились.

КАК ЯЗЫКОВЫЕ МОДЕЛИ ИЗМЕНИЛИ НАШУ ЖИЗНЬ

Появление моделей, основанных на технологиях обработки естественного языка (большие языковые модели, **Large Language Model** — LLM), радикально изменило общение людей с окружающим миром. Мы получили возможность не только мгновенно получать ответы на сложные вопросы и создавать программы, не зная, как писать код, но и вступать в диалог с машиной, а также использовать ИИ для общения с другими людьми.

Когда появился Chat GPT, пользователи Tinder начали активно применять его для общения с потенциальными партнерами.

ЭВОЛЮЦИЯ ЯЗЫКОВЫХ МОДЕЛЕЙ

Системы NLP (Natural Language Processing), разработка которых началась еще в 1950-х годах, можно разделить на два типа:

1. первые работают по заранее заданным параметрам и созданным человеком правилам;
2. вторые — на основе машинного обучения, самостоятельно формируя правила из полученных данных.

Ранние системы обработки языка использовали условную логику: если встречается определенное слово, система дает заранее подготовленный ответ. Такие системы было сложно масштабировать и учить отвечать на сложные запросы.

С конца 1990-х годов разработчики начали фокусироваться на создании узкоспециализированных чат-ботов, которые могли поддерживать диалог в пределах конкретной темы. Для улучшения взаимодействия стали использовать сопоставление шаблонов: вместо прописывания всех возможных фраз создавались универсальные конструкции, позволяющие распознавать общие смыслы в репликах пользователей.

В 2000-е годы активно развивался анализ настроений: системы учились интерпретировать эмоции пользователя, например использовать капслок или ругательства как индикаторы гнева. Это должно было сделать взаимодействие более «человечным», но часто приводило к неестественным и фальшивым ответам. Параллельно добавлялись антропоморфные элементы: чат-ботам давали имена, их обучали реагировать на эмоциональные фразы и выстраивать ответы с намеком на сочувствие или юмор. Все это усиливало иллюзию общения с живым собеседником.

В первом десятилетии XXI века началось внедрение более продвинутых методов предобработки текста, среди них — автоматическое исправление орфографических ошибок, нормализация диалектных форм и стемминг (выделение основы слова для облегчения анализа и сопоставления смыслов). Эти приемы позволяли системам лучше понимать разнообразные пользовательские формулировки. Если запрос оказывался непонятным, система переходила к заранее заготовленным универсальным ответам. Появились и первые попытки использовать память: краткосрочную — для учета контекста текущего диалога и долгосрочную — для запоминания предпочтений пользователя между сессиями. Это создавало ощущение более индивидуального (и человеческого) общения.

Наконец, в 2010-х появились статистические языковые модели (SLM), которые обучались на больших объемах текстов и предсказывали наиболее вероятное продолжение фразы на основе частотности словосочетаний. А затем — и нейросетевые модели, такие как LSTM (Long Short-Term Memory), которые позволили учитывать более длинный контекст и последовательно обрабатывать текст с сохранением связи между частями фразы.

Статистические и нейросетевые языковые модели активно применяются в:

- распознавании речи — для повышения точности за счет учета контекста;
- машинном переводе — для анализа вероятностей соответствия слов в разных языках;
- предсказании текста — в пользовательских интерфейсах, когда система помогает продолжить предложение;
- автоматической вычитке — при анализе и корректировке больших объемов текста.

В 2014 году компания Amazon выпустила голосового ассистента Alexa вместе с первым устройством Amazon Echo. В 2016-м на конференции Google I/O был представлен Google Assistant, который сначала был встроен в мессенджер Allo и устройство Google Home, затем появился на смартфонах Pixel, а позже — на большинстве Android-устройств. Обе системы использовали достижения в области NLP, голосового распознавания и машинного обучения, делая технологии разговорного ИИ частью повседневной жизни.

В 2017 году была впервые описана архитектура трансформеров — принципиально новый подход к языковому моделированию, позволивший моделям обрабатывать текст не последовательно, а с помощью механизма

В диагностической системе для выявления инфекций крови MYCIN, созданной в 1970-х годах, использовалось 500 логических правил. Она была успешна, но требовала очень больших вычислительных мощностей, что затрудняло ее широкое внедрение.

внимания (attention), одновременно учитывая множество данных. На основе этой архитектуры были созданы мощные языковые модели, такие как GPT, способные анализировать и генерировать тексты, вести диалог, отвечать на сложные вопросы и решать задачи, которые ранее считались недоступными для машинного интеллекта.

В 2020 году Алексей Досовицкий предложил Vision Transformer, что значительно улучшило распознавание изображений. С 2022 года трансформеры начали применяться для обработки аудио и для выполнения задач машинного обучения.

ИЛЛЮЗИЯ ПОНИМАНИЯ

Проблема сходства и различия между человеческим и машинным интеллектом была поставлена еще в середине XX века, когда развитие технологий обработки естественного языка только начиналось.

В 1950 году британский математик и криптограф Алан Тьюринг предложил тест, в котором участник должен был определить, кто из его собеседников человек, а кто — машина. Если компьютер успешно выдавал себя за человека, тест Тьюринга считался пройденным.

В 1980 году американский философ Джон Сёрль взялся доказать, что машина может пройти тест Тьюринга, то есть убедительно общаться, даже не понимая сути происходящего. Сёрль предложил мысленный эксперимент под названием «Китайская комната», в котором человек, не знающий китайского языка, подбирает иероглифы по постоянно совершенствуемым инструкциям и выдает осмысленные ответы на заданные вопросы. «Китайская комната» показала, что ИИ создает иллюзию осознанности, следуя алгоритмам, но не имея реального понимания.

Три года спустя, в 1983 году, появилась первая генеративная текстовая программа Racter, которая создавала прозаические и поэтические тексты на английском языке на основе заложенных правил, языковых шаблонов и словарей. Написанная Racter книга The Policeman's Beard Is Half Constructed была опубликована в 1984 году и показала, где проходит граница между текстом как формой и текстом как содержанием. Текст, созданный Racter, был воспринят многими как бессвязный, абсурдный и сюрреалистичный.

Британский профессор когнитивной робототехники Мюррей Шанахан подчеркивает, что задача LLM — предсказание текстовых последовательностей, что позволяет решать различные задачи — от ответов на вопросы до перевода — без понимания сути процесса.

Современные большие языковые модели (LLM) «учатся» на огромном количестве текстов и могут вводить в заблуждение даже экспертов, создавая впечатление осознанности. Они создают иллюзию осведомленности, социального интеллекта и эмоций, но на самом деле просто генерируют текст, опираясь на языковые паттерны. Искусственный интеллект, даже самый продвинутый, не испытывает эмоций и не умеет сочувствовать по-настоящему. Он просто подбирает наиболее вероятные ответы, основываясь на том, что видел в текстах.

Ключевое различие между машиной и человеком — социальный интеллект, то есть способность понимать и адекватно отражать смысловой и эмоциональный контекст.

По мере роста мощностей у языковых моделей появляются способности, не прописанные заранее. Эти так называемые эмерджентные свойства включают решение логических задач и генерацию убедительной речи, а также способность анализировать психологическое состояние собеседника. Метод обучения на основе человеческой обратной связи (RLHF) позволяет моделям все точнее отражать человеческие ценности и предпочтения.

Современные ИИ-системы уже умеют шутить и выражать сочувствие, поэтому исследователи считают, что развитие языковых моделей может стать ключом к созданию общего искусственного интеллекта (AGI).

Чем больше взаимодействие с ИИ похоже на общение с человеком, тем уязвимее мы становимся. Доверие растет, а то, что у ИИ нет моральных принципов, — забывается.

Искусственный интеллект уже сегодня массово применяется в социальных манипуляциях, политической пропаганде, слежке и прямом мошенничестве, а также в создании современного оружия.

Вопрос состоит не в том, будет ли ИИ использоваться как оружие, а в том, против кого это оружие будет направлено: будут ли люди использовать ИИ для своих целей, или ИИ будет манипулировать людьми.

1 сентября 2017 года на встрече со студентами в День знаний Президент России Владимир Путин отметил, что ИИ станет ключевым фактором в глобальном распределении сил.

ИСКУССТВО ОБМАНА: СОЦИАЛЬНАЯ ИНЖЕНЕРИЯ ВЧЕРА И СЕГОДНЯ

Социальная инженерия, то есть манипуляции с целью воздействия на поведение человека, существует столько же, сколько само человечество. Примеры обмана и манипуляций можно найти в сказках, мифах и баснях всех народов, а также, конечно, в современной политике, экономике и управлении.

Но сам термин «социальная инженерия», который получил широкую известность в конце XX — начале XXI века, появился в 1894 году. Его автор, нидерландский предприниматель, специалист в области промышленного менеджмента Ян Якоб ван Маркен, подчеркивал необходимость учитывать не только технические, но и человеческие аспекты при организации производства и внедрении технологических новшеств.

Сегодня приемы социальной инженерии широко используются киберпреступниками для обхода технической защиты. Злоумышленники используют фишинг, поддельные письма и вредоносные файлы, чтобы заставить человека создать лазейки для проникновения в корпоративные, банковские и другие системы.

В книге Роберта Чалдини «Психология убеждения» описаны шесть ключевых факторов, которые позволяют успешно манипулировать поведением людей:

1) **взаимный обмен** — стремление отплатить за оказанную услугу;
2) **дефицит** — создание ощущения, что возможность редкая или скоро исчезнет;
3) **авторитет** — подчинение людям, которые кажутся компетентными;
4) **консенсус** — следование за большинством в неопределенности;
5) **последовательность** — действия в соответствии с предыдущими решениями;
6) **симпатия** — доверие к тем, кто вызывает положительные эмоции.

Мошенник Фрэнк Абигнейл (фильм «Поймай меня, если сможешь»), обманывал людей, надевая форму пилота или медицинский халат.

Принципы взаимности и авторитета не очень трудоемки в применении и могут быть использованы в короткие сроки.

Принцип симпатии — один из самых эффективных методов социальной инженерии, но требует времени и усилий для создания доверительных отношений. На этот принцип опираются агенты спецслужб, годами добывающие секретную информацию.

В конце 1970-х КГБ направил агента Альбрехта Дитриха, действовавшего под именем Джек Барски, в США с целью внедрения в американское общество и сбора разведывательной информации. Дитрих в течение 10 лет жил под вымышленной личностью, работал в сфере IT и устанавливал контакты в политических кругах. Этот случай стал одним из ярких примеров использования социальной инженерии в шпионской практике.

В фильме Ex Machina робот Ава создает иллюзию эмоций и привязанности, манипулирует человеком и в конце концов достигает своих целей, действуя безжалостно и изобретательно.

Автор подчеркивает, что использование психологических принципов в исполнении ИИ становится еще опаснее, потому что ИИ не имеет никаких моральных ограничений.

Приемы социального инжиниринга сегодня применяются в таких видах воздействия, как:

- спам: объемы нежелательной рекламы остаются огромными, несмотря на законы против спама. Мошенники используют современные технологии для сбора адресов электронной почты, массовой рассылки фишинга и вирусов;
- точечный маркетинг («капитализм слежки»): с помощью таргетированной рекламы и машинного обучения компании могут влиять на поведение пользователей, используя их личные данные;
- цензура и пропаганда в интернете;
- боты — автоматические программы, которые распространяют ложную информацию, манипулируют, выманивают личные и платежные данные и даже шантажируют людей. Часто жертва при этом не осознает, что общается с ИИ. А приемы психологической манипуляции приводят к тому, что на удочку преступников попадаются даже те, кто знает о возможных рисках.

Продвинутые языковые модели дают возможность указания ролей и целей бота через API (Application Programming Interface), что делает их особенно удобными для создания автономных вредоносных ботов.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ИЛИ ИСКУССТВЕННОЕ СОЗНАНИЕ?

Сегодня многие думают, что ИИ скоро станет по-настоящему разумным, самостоятельным и даже способным чувствовать и повернуть свою мощь против человечества.

Однако в основе этих страхов часто лежит непонимание того, как работает технология.

Осознанность и способность чувствовать — это субъективные процессы, которые связаны с восприятием ощущений и мышлением. Несмотря на то что современные языковые модели обладают огромными вычислительными способностями и могут выполнять некоторые задачи лучше человека, они:

- не воспринимают выполнение задачи как осознанный процесс;
- не умеют быть автономными, то есть действовать на основе собственных решений.

Нынешние технологии далеки от создания осознанного и самостоятельного искусственного интеллекта, хотя в будущем это теоретически возможно.

ТЕХНИЧЕСКИЕ УЯЗВИМОСТИ ИИ

В последние годы большие языковые модели (LLM) привлекают внимание и в контексте технических угроз. Эти системы могут взаимодействовать с компьютерными системами, работать с данными в форматах CSV, JSON, XML и создавать код на различных языках программирования.

Хатчинс пишет, что с LLM связаны многие риски:

1. **Неточная интерпретация языка.** Ошибки при преобразовании человеческих инструкций в машинные действия могут приводить к некорректному выполнению задач.

2. **Ограниченная управляемость.** Модели чувствительны к формулировке запросов и могут демонстрировать нежелательное поведение при неясных или двусмысленных инструкциях.
3. **Сложности в диагностике.** Отсутствие привычного исходного кода затрудняет анализ внутренних процессов и поиск причин сбоев.
4. **Непредсказуемость поведения.** Модели могут давать неожиданные ответы в новых или нестандартных ситуациях, что ограничивает их надежность.
5. **Трудности корректировки.** В отличие от традиционного ПО, языковые модели сложно донастраивать вручную для исправления конкретных ошибок.
6. **Уязвимость к злоупотреблениям:** языковые модели имеют скрытые слабые места, которые потенциально можно использовать для нанесения ущерба или обхода ограничений.
7. **Злоумышленное техническое применение** LLM для автоматизации атак на компьютерные системы, включая поиск и эксплуатацию уязвимостей.
8. **Создание автономных вредоносных систем** — построение хакерских ИИ, способных самостоятельно инициировать и развивать атаки, без участия человека
9. **Мультимодальная уязвимость,** то есть распространение атак на создание изображений, аудио, видео и других типов данных.
10. **Интеграционные риски:** объединение различных типов данных в единой системе может приводить к непредсказуемым сценариям использования и усилению манипулятивного потенциала.

Пример мультимодальной уязвимости — распространение deepfake-контента — фальшивых изображений, аудио и видео, созданных с помощью ИИ и неотличимых от реальных. Так, в 2023 году в Сети появились фальшивые изображения ареста Дональда Трампа, созданные нейросетью. Они быстро разошлись по соцсетям и были использованы для дезинформации. Такие подделки подрывают доверие к реальным событиям и могут спровоцировать политическую нестабильность.

БУДУЩЕЕ ИИ: РОБОТЫ, НЕЙРОИНТЕРФЕЙСЫ И АВТОНОМНОЕ ОРУЖИЕ

Мы живем в эпоху, когда идеи, еще недавно казавшиеся фантастикой, становятся реальностью, открывая новые возможности и создавая новые риски.

Одно из важных направлений развития ИИ — это его интеграция с физическими роботами. Трансформеры позволяют легко связать механические действия с человеческим языком. Это может привести к массовому распространению роботизированных систем и человекоподобных роботов под управлением ИИ. Однако такая интеграция несет серьезные риски, прежде всего связанные с изменениями программного обеспечения.

Продолжаются разработки интерфейсов «мозг-компьютер», которые позволят напрямую подключать человеческий мозг к вычислительным системам. Это создает риски утраты конфиденциальности и контроля, а также может привести к размыванию границ человеческой личности.

Системы, которые могут принимать решения сами, рискуют вызвать «молниеносные войны», где технологии будут действовать быстрее, чем люди смогут реагировать. Это создает серьезные угрозы для мировой безопасности.

Уже сегодня лицо войны меняется из-за массового применения беспилотных летательных аппаратов. Автор предупреждает о новой гонке вооружений в сфере автономного оружия.

С развитием генеративных систем все больше контента в интернете будет создавать ИИ. Это подрывает сам принцип доверия к информации. А поскольку ИИ обучается на созданных им же данных, минимальная предвзятость, заложенная на начальном этапе, может нарастать, искажения данных — усиливаться, что может привести к «краху моделей».

КАК УПРАВЛЯТЬ ИИ: СОТРУДНИЧЕСТВО, ПРОСВЕЩЕНИЕ, РЕГУЛИРОВАНИЕ

В 2005 году в книге The Singularity Is Near американский изобретатель и футуролог Рэй Курцвейл предсказал, что примерно в 2045 году наступит сингулярность — момент, когда ИИ станет умнее человека и кардинально изменит цивилизацию.

Мы приближаемся к этому моменту, и его последствия трудно предсказать. Пока мы не знаем, как человечество будет взаимодействовать со сверхразумными системами и как мы сможем сохранить свою значимость в мире, где машины будут умнее людей.

Однако уже сегодня понятно, что для эффективного регулирования и безопасного использования ИИ необходимо тесное международное сотрудничество, ради которого нужно научиться преодолевать многочисленные разногласия. Ни одна страна или организация не сможет в одиночку решить проблемы, связанные с развитием ИИ.

Хатчинс называет несколько направлений, в которых необходимы немедленные действия:

1. **борьба с дезинформацией и манипуляциями** с помощью ИИ-моделей, способных отличать сгенерированный контент от созданного человеком;
2. **повышение общественной грамотности в сфере кибербезопасности,** развитие навыков применения многоуровневых инструментов (таких как двухфакторная идентификация);
3. **внедрение практик управления рисками на этапе разработки ИИ,** установление четких правил и жесткого контроля за их исполнением в компаниях государственного и коммерческого секторов.

---

10 ЛУЧШИХ МЫСЛЕЙ

1.

Развитие моделей анализа естественного языка шло от систем с жесткими наборами правил в сторону большей гибкости и прогнозирования пользовательского поведения.

2.

Современные языковые модели ИИ основаны на архитектуре трансформеров, которая позволяет анализировать текст с учетом контекста сразу в нескольких направлениях.

3.

Модели-трансформеры учатся на получаемых ими данных и не имеют четкого программного кода, что дает им огромную гибкость, но может приводить к нарастанию ошибок.

4.

Современные инструменты ИИ могут практически идеально имитировать человеческое общение и генерировать аудио- и видеоконтент, который очень сложно отличить от настоящего.

5.

Чем успешнее ИИ создает иллюзию живого общения, тем больше растет уязвимость людей перед манипуляциями в цифровой среде.

6.

Принципы манипуляции поведением людей (взаимность, дефицит, авторитет, консенсус, последовательность, симпатия) не меняются веками.

7.

Социальная инженерия, применяемая с помощью ИИ, особенно опасна, так как у искусственного интеллекта нет моральных ограничений.

8.

Множество страхов, связанных с ИИ, необоснованны с точки зрения технологий. Но ИИ уже сегодня широко применяется в целях пропаганды, мошенничества, слежки, разведки и ведения реальных кровопролитных войн.

9.

Технические уязвимости, связанные с использованием ИИ, могут стать причиной критических сбоев в системах, приводя к утрате данных, финансовым потерям, нарушению работы инфраструктуры и созданию угроз для безопасности людей.

10.

Эффективная защита от угроз, связанных с развитием искусственного интеллекта, требует международного сотрудничества, просвещения пользователей и жесткого регулирования требований к компаниям-разработчикам ИИ.